

## Identifying novel protein interactions

Gonçalves Carneiro, Vítor; Clarke, Thomas; Davies, Clare C; Bailey, Dalan

DOI:

[10.1016/j.ymeth.2015.08.022](https://doi.org/10.1016/j.ymeth.2015.08.022)

License:

Creative Commons: Attribution-NonCommercial-NoDerivs (CC BY-NC-ND)

*Document Version*

Peer reviewed version

*Citation for published version (Harvard):*

Gonçalves Carneiro, V, Clarke, T, Davies, CC & Bailey, D 2015, 'Identifying novel protein interactions: proteomic methods, optimisation approaches and data analysis pipelines', *Methods*.  
<https://doi.org/10.1016/j.ymeth.2015.08.022>

[Link to publication on Research at Birmingham portal](#)

### **Publisher Rights Statement:**

Eligibility for repository: Checked on 29/10/2015

### **General rights**

Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

- Users may freely distribute the URL that is used to identify this publication.
- Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
- User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
- Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

### **Take down policy**

While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact [UBIRA@lists.bham.ac.uk](mailto:UBIRA@lists.bham.ac.uk) providing details and we will remove access to the work immediately and investigate.

## Accepted Manuscript

Identifying novel protein interactions: proteomic methods, optimisation approaches and data analysis pipelines

Daniel Gonçalves Carneiro, Thomas Clarke, Clare C. Davies, Dalan Bailey

PII: S1046-2023(15)30066-9

DOI: <http://dx.doi.org/10.1016/j.ymeth.2015.08.022>

Reference: YMETH 3779

To appear in: *Methods*

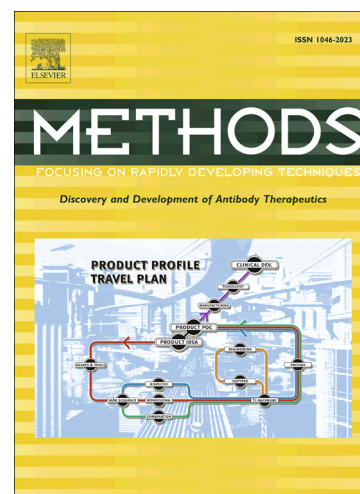
Received Date: 30 June 2015

Revised Date: 26 August 2015

Accepted Date: 27 August 2015

Please cite this article as: D.G. Carneiro, T. Clarke, C.C. Davies, D. Bailey, Identifying novel protein interactions: proteomic methods, optimisation approaches and data analysis pipelines, *Methods* (2015), doi: <http://dx.doi.org/10.1016/j.ymeth.2015.08.022>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



# Identifying novel protein interactions: proteomic methods, optimisation approaches and data analysis pipelines

Daniel Gonçalves Carneiro<sup>1</sup>, Thomas Clarke<sup>1</sup>, Clare C Davies<sup>\*1</sup> and Dalan Bailey<sup>\*1</sup>

<sup>1</sup> College of Medical and Dental Sciences  
University of Birmingham  
Edgbaston  
Birmingham  
B15 2TT  
UK

\* Co-corresponding authors: email [d.bailey@bham.ac.uk](mailto:d.bailey@bham.ac.uk); [c.c.davies@bham.ac.uk](mailto:c.c.davies@bham.ac.uk)

Telephone: +441214146854

## Abstract:

The technological revolution in high-throughput nucleic acid and protein analysis in the last 15 years has launched the field of 'omics and led to great advances in our understanding of cell biology. Consequently the study of the cellular proteome and protein dynamics, in particular interactomics, has been a matter of intense investigation, specifically the determination and description of complex protein interaction networks in the cell, not only with other proteins but also with RNA and DNA. The analysis of these interactions, beginning with their identification and ultimately resulting in structural level examination, is one of the cornerstones of modern biological science underpinning basic research and impacting on applied biology, biomedicine and drug discovery. In this review we summarise a selection of emerging and established techniques currently being applied in this field with a particular focus on affinity-based purification systems and their optimisation, including tandem affinity purification (TAP) tagging, isolation of proteins on nascent DNA (IPOND) and RNA-Protein immunoprecipitation in tandem (RIPiT). The recent application of quantitative proteomics to improve stringency and specificity is also discussed, including the use of metabolic labelling by stable isotope labelling by amino acids in cell culture (SILAC), localization of organelle proteins by isotope tagging (LOPIT) and proximity-dependent biotin identification (BioID). Finally, we describe a range of software resources that can be applied to interactomics, both to handle raw data and also to scrutinise its broader biological context. In this section we focus especially on open-access online interactomic databases such as Reactome and IntAct.

**Keywords:** interactomics, protein-protein interaction, RNA-protein interaction, DNA-protein interactions, affinity purification, TAP-tagging

## Introduction:

Fundamentally, cellular life depends on a continuing network of molecular interactions among proteins, nucleic acids and small molecules. Interactions of this nature underpin the key processes of transcription and translation and are essential for supporting cellular structure, metabolism, cell signalling and all other aspects of cell biology. Aberrant interactions, caused by genetic or environmental stimuli, are the cause of most human diseases such as cancer or auto-immunity, while all pathogens, in one way or another, hijack these interactions to promote their own growth over that of the host. As a result the description of these interactions is, to some degree, the central principle of all molecular biology research. Their isolation, identification and characterisation can be either straightforward or complex; this tenet being dependent on the very nature of the interaction itself.

For the purposes of this review we have focused on examining techniques designed to identify three of the major types of interactions found in the cell: Protein-protein, DNA-protein and RNA-protein. This obviously excludes other interactions such as small molecules and proteins; however, these have been excellently addressed elsewhere [1-3]. In this article we will particularly focus on recent advancements in these fields, considerations that should be made on the choice of methodology as well as potential limitations to bear in mind when designing experiments (overview; Figure 1). We will also provide a basic overview of these techniques and their application as a starting point for researchers, e.g. structural biologists, looking to implement these technologies in their research along with a list of resources to facilitate downstream data processing.

## Methods For Identifying Protein-Protein Interactions:

The two commonly used approaches for discovery of protein-protein interactions are: (1) the use of recombinant protein libraries that can be screened with a bait protein or (2) the isolation and *de novo* detection of protein complexes using mass spectrometry (MS).

**Yeast-Two-Hybrid:** Pioneering interactomic techniques, such as yeast two-hybrid (Y2H), were important initial tools for dissecting protein-protein interactions in cells [4]. In this system, two recombinant fusion proteins are generated: the bait is fused with a DNA-binding domain and the prey with a transcription activation domain. These two elements are expressed in yeast along with a reporter gene whose regulation is under a promoter recognised by the recombinant transcription activation domain. If these two proteins interact with each other, the transcription activation domain becomes functional leading to expression of the reporter gene. Despite being a robust assay, this system is limited by the preference towards high affinity interactions that happen in the nucleus, by protein complexes that form in cellular membranes or cases where proteins require more than one binding partner [5]. Recent advances in the field have led to the development of variations of Y2H,

specialising in the recruitment of the bait and prey to the plasma membrane [6], cytosol [7, 8] or the endoplasmic reticulum [9], and also expanding the number of proteins included, such as multiple baits [10]. Nonetheless, this technique is still largely limited by the choice of library. Identifying protein-protein interactions relies on specific properties of proteins beyond their peptide sequence, such as protein abundance, cellular localization, post-translational modifications and isoforms or variants; most of which can be missed by surrogate systems as Y2H [11]. Some of these limitations have been addressed with the development of more powerful hybrid technologies that can successfully isolate macromolecular protein assemblies, i.e. by using split-ubiquitin membrane yeast two-hybrid (MYTH) where bait and prey proteins are tagged with split ubiquitin [12]. Specific protein-protein interactions result in reconstitution of the ubiquitin and cleavage – a process that releases a fused transcription factor and leads to reporter protein activation. This approach has proved particularly successful for isolating membrane protein interactions since normal Y2H targeting of hydrophobic proteins to the nucleoplasm can be inefficient. Subsequent improvements have included the use of genomically integrated baits (iMYTH) [13] to provide more endogenous protein levels as well as adaptation of the system to mammalian cells (MaMTH) [14].

**Protein microarrays:** The advent of high-throughput technology has allowed the rapid screening of protein interactions in different organisms. Protein microarrays, in particular, are increasingly being used to study protein interactions with applications being extended to diagnostic [15] or enzymatic discovery [16]. In a protein microarray, which is a derivation of the classical DNA chip for detection of RNA, a library of complete or functional domains of recombinant proteins is printed onto a slide [17]. A tagged bait protein is then passed over the slide allowing it to specifically interact with a relevant prey protein, before detection by tag-specific immunoblotting. Protein microarrays have allowed the determination of not only protein-protein interactions in the context of diverse network modalities and functional modifications (reviewed in [18]) but also RNA-protein [19] and DNA-protein interactions [20, 21]. Although an increasing diversity of microarray chips are commercially available at competitive prices, these methods are limited by the necessity to reproduce a putative protein interaction *in vitro* and the recapitulation of the physiochemical and stoichiometric environment necessary for the interaction.

**Mass spectrometry-based approaches:** Most affinity purification techniques rely on the isolation of material for subsequent MS-based protein identification. In this process, a complex protein sample is separated by size and isoelectric point - both 2D gel electrophoresis and liquid chromatography (LC) are common practices in proteomics. This is followed by proteolysis often through tryptic digestion, a process that generates small peptides which are subsequently exposed to ionizing radiation and analysed by a mass spectrometer. During this procedure each peptide is broken into smaller ions whose mass-to-charge ( $m/z$ ) ratios are used to extrapolate the peptide

sequence [22]. The identification and quantification of these peptides, achieved through computational comparison with protein databases, allows the reconstruction of a putative protein profile for a sample. However, the identification of a specific protein can lead to false-positives due to uncategorised isoforms or protein variants (since a peptide can originate from more than one protein) [23, 24]. The use of LC-MS to study protein-protein interactions has two main limitations: (1) the relative quantification and identification of specific versus non-specific interactions and (2) the overall complexity of the protein profile in the sample. Other limitations do also exist: the identification of low-yield proteins, the transient state of protein interactions, and the requirement for co-factors and/or post-translation modifications; however, these are often related to specific interactions and will not be addressed in any detail in this review. Of note, with regards to identifying low-yield proteins, the improved sensitivity of modern MS technology is somewhat rectifying these issues, as is the availability of improved affinity purification reagents and techniques. Nevertheless, care should always be taken when designing interactomics experiments to consider the amount of starting material required, often cell number.

**Immunoprecipitation techniques:** The application of new protein-protein interaction techniques combined with affordable LC and MS approaches has revolutionized interactomics research. The most commonly used technique for isolating protein interactors is affinity purification, usually employing ectopically-expressed tagged (e.g. GFP, HA, Flag) bait proteins. This process has advantages over Y2H in that it allows the efficient study of proteins in more physiological situations such as mammalian cells. In this methodology, the protein of interest and any complex it forms in the cell are recovered by the use of high affinity reagents (e.g. antibodies or peptides coupled to agarose beads) to the tag. This reliable and flexible approach has been applied in a range of biological contexts to study both small protein complexes as well as global protein-protein interaction networks [25, 26]. The clear strength of this method is that it allows for the enrichment of final protein complexes that occur less frequently or have low abundance in the cell; however, optimization of the purification protocol is essential to simultaneously maintain the protein complexes but also to eliminate possible contamination through non-specific protein interactions with the bait, tag or affinity matrix.

**Tandem Affinity Purification:** Issues with interaction stringency during immunoprecipitation can be resolved using tandem affinity purification (TAP) [27]. This technology relies on the application of dual affinity tagged-fusion proteins containing a strong antigenic region (such as Protein G, GFP, etc.) fused to a separate tag, such as streptavidin or calmodulin-binding peptide. The two tags are normally separated by a specific exogenous cleavage site, such as the Tobacco Etch Virus or PreScission (GE Healthcare) or Rhinovirus 3C protease cleavage sites. After expression of a TAP-tagged bait, protein complexes are isolated by two separate affinity purifications. The cleavable linkers and small peptide tags used in the second purification step allow specific elutions to be

performed after each purification further increasing the specificity of the pull-downs, reviewed in more detail here [28, 29]. This technique has been used to study protein-protein interactions in a variety of fields, such as virology, oncology and cell biology and is one of the most popular methods for identifying new protein binding partners [30-33].

Despite its broad application, large-scale analysis of the interactome using TAP tagging is time consuming and expensive. In addition, due to the thorough purification technique, weaker or transient binding partners may not be detected in the final protein preparation. Moreover, the use of recombinant proteins, generally overexpressed from complementary DNA, can affect localization, endogenous abundance and hence protein complex stoichiometry, turnover and the general function of the protein of interest, ultimately influencing the nature of the protein-protein interaction. It is therefore essential that any protein interaction identified should be subsequently confirmed under more physiological conditions, i.e. with endogenous untagged protein. One emerging technique to counter some of these issues is to normalise the expression of the tagged protein to a more physiologically relevant level. Bacterial artificial chromosomes (BACs), large enough to contain the gene of interest as well as its endogenous regulatory sequences, have recently been applied to assist this normalisation. In fact, BACs libraries fused with TAP tags have facilitated high-throughput interactomic studies of different cells [34, 35].

#### **Methods For Identifying DNA-Protein Interactions:**

The identification of DNA-protein interactions underpins basic research in a number of crucial biological fields including DNA replication, DNA repair and chromatin remodelling. As such there is a great interest in developing new techniques to examine this in as close to a physiological setting as possible.

**IPOND:** There are a range of emerging techniques to identify proteins that bind DNA; one popular method in recent years has been isolation of proteins on nascent DNA (IPOND). This technique utilises the thymidine analogue 5-ethynyl-2'-deoxyuridine (EdU) which, when added to cells growing in tissue culture, is rapidly incorporated into newly synthesised DNA [36]. Crucially, this occurs without interfering with the replication machinery or causing any measurable activation of the cell's DNA damage response and is therefore not considered deleterious to cell viability and function [37]. EdU labelled cells can then be treated with a reversible cross-linking agent to facilitate the isolation of any DNA-interacting proteins [38]. The functional alkyne group in EdU also allows the addition of a biotin tag (through click chemistry) to the nascent DNA which can facilitate purification and the subsequent isolation of protein interactors by immunoblotting or MS.

***Xenopus laevis* eggs:** Another increasingly popular method, particularly in the study of DNA replication, is the application and manipulation of *Xenopus laevis* eggs for DNA-protein



interactomics. These soluble egg extracts are an attractive model as they undergo the same nuclear transitions of the eukaryotic cell cycle *in vitro* as they do naturally *in vivo* [39]. *Xenopus* eggs are typically around 1mm in diameter and contain a vast array of RNAs and proteins. These are required to synthesise up to 4000 nuclei as a result of 12 synchronous rounds of cell cycle; a process which occurs immediately after fertilisation in the absence of subsequent rounds of transcription [40]. Typically, before fertilisation occurs, *Xenopus* egg extracts are biologically arrested in metaphase of meiosis II [41]. Exogenous DNA added to the egg extracts is first organised into chromatin and then into structures that correspond to those of cells in interphase, thereby enabling DNA duplication to occur [39]. Given that the *Xenopus* egg extract is able to support such a wide range of biological activity, ranging from *in vitro* DNA replication to mitosis, they represent a unique tool for interrogating the dynamics and mechanisms that underpin such processes. In addition, the egg extracts form a soluble soup-like mixture that is amenable to immuno-depletion or supplementation with recombinant protein. Likewise, small molecule inhibitors to various enzymes or DNA damaging agents can be easily added. This further enhances their suitability as a tool to identify and investigate DNA-protein interactions. These techniques have been applied to isolate chromatin and chromatin-associated proteins from egg extracts with DNA-protein interactions subsequently being identified by MS [42] or immunoblotting [43]. In addition, Moreno *et al.*, have recently utilised this method to study the role of polyubiquitylation in the termination of DNA replication, demonstrating that inhibition of polyubiquitylation caused a prolonged association of the active helicase with replicating chromatin [44]. As such these methods can answer key questions relating not only to DNA and protein interactions but also to the duration of these interactions and their significance for biological processes.

**DNA baits:** Using DNA directly as the bait for identifying DNA-protein interactions is also a popular technique. One such application is the identification of guanine quadruplex (G-quadruplex) binding proteins. G-quadruplexes are unusual secondary structures which typically occur as a result of guanine rich regions within DNA or RNA [45]. Their recent identification in the promoter regions of genes implicated in cancer suggests a regulatory role in transcription [46]. Binding proteins have been identified using pull down assays where G-quadruplexes serve as the bait. In this instance specificity can be determined using mutated oligonucleotides which are unable to form defined G-quadruplex structures. In these experiments, which aim to examine how these complex structures affect cellular processes such as transcription and translation, isolated proteins are identified by MS [47].

#### **Methods For Identifying RNA-Protein Interactions:**

RNA processing, including alternative mRNA splicing and editing, is central for the increased proteome complexity associated with higher eukaryotes. Consequently, more than 1000 RNA-



binding proteins (RBPs) are encoded by the genome. However, with the explosion of technological advances enabling whole genome transcriptional analysis, it is becoming increasingly clear that RNA is not solely an intermediate between DNA and protein production. Indeed, whilst only 2% of our genome encodes for proteins, up to 70% is transcribed into non-coding RNA, including long non-coding RNAs (lncRNAs). Both mutations in RBPs and aberrant expression of lncRNAs have been implicated in human disease, such as neurodegeneration and cancer [48]. Despite this, little is known about how RBPs control cell-type specific alternative splicing of mRNAs and lncRNA biology. Interaction studies and structural insights will undoubtedly be central in providing a deeper understanding of these processes. The “omics” era has led to a rapid expansion in methodology to study the RNA-protein landscape. Traditionally, two main approaches are employed: protein-centred, whereby specific proteins serve as a bait to isolate associated RNA which is identified by sequencing; or RNA-centred involving the expression of affinity-coupled RNA and MS to identify proteins that bind a specific RNA species.

**Protein-centred approaches:** Chromatin immunoprecipitation (ChIP) is a well-established technique that enables the isolation of protein:DNA complexes via antibody-mediated purification of endogenous or tagged proteins. RNA immunoprecipitation (RIP) is analogous to chromatin immunoprecipitation (ChIP), but in this case, RBPs are affinity purified and their associated RNAs isolated. RNA is then converted to cDNA and target sequences identified by quantitative PCR, cDNA libraries, microarray analysis or RNA-Seq [49]. Unfortunately, RBPs are relatively promiscuous when overexpressed, thus it is important to generate stable cell lines that express levels close to that of endogenous protein. Accordingly, CRISPR technology enabling the introduction of tag sequences within gene sequences upstream of the open reading frame will undoubtedly facilitate this. Of note, His-tagged sequences are not recommended as the positive charge can alter the RNA binding properties of the RBP.

Like ChIP, RIP can be performed under native or denaturing conditions; however, whilst native RIP potentially offers physiological isolation of complexes, extensive non-specific interactions that can occur between isolated RBPs and RNA after cell lysis can be problematic [50]. To circumvent this, RNA-protein complexes are cross-linked by treating cells with ultraviolet C (UVC) light, enabling denaturation of samples and stringent lysis conditions (cross-linking and immunoprecipitation, CLIP) [51]. A further development on this strategy is photoactivatable ribonucleoside-enhanced (PAR)-CLIP, enabling single-nucleotide resolution of protein interaction sites and the definition of sequence motifs [52]. However, UVC-induced cross-linking is far more efficient when the RNA-protein interface is composed of nucleic acid bases and aromatic amino acid side chains, which is often but not always the case for RBPs [53]. CLIP techniques are therefore inherently biased towards sites that are more effectively cross-linked, and thus may not be a true representation of the RNA-binding landscape. To overcome this, a combinatorial method of RIP and CLIP has recently been devised,

RIPIT (RNA-Protein immunoprecipitation in tandem) [54], that involves two sequential immunoprecipitations, and is thus highly analogous to the isolation of protein-protein complexes by TAP-tagging. Importantly, RNA-protein complexes are cross-linked with formaldehyde rather than UVC and the noise-to-signal ratio of RNA-binding footprints is greatly enhanced by the two-stage purification process. Indeed, whilst direct RNA-protein interactions would need verification, the ability to perform sequential immunoprecipitations of two different proteins within an RNA-bound complex will enable detailed and dynamic analysis of multi-subunit RBPs.

**RNA-centred approaches:** A reverse but complementary approach to study ribonucleoprotein (RNP) complexes involves identifying trans-regulatory proteins that interact with a specific RNA sequence. Since RNP complexes are highly dynamic and multifaceted, a robust *in vivo* system is required. The “Ribotrap” method, developed by Keene and colleagues [55], was one of the first to exploit the phenomenon of naturally occurring RNA-RBP protein interactions, such as the bacteriophage MS2 viral coat protein that recognises with high specificity an MS2-RNA stem loop structure. By engineering and expressing two constructs, one that will express the MS2-RNA stem loop conjugated to the RNA bait and the other that will express the MS2 protein conjugated to an affinity tag (e.g. GST, protein A), *in vivo* RNA-RBP complexes can be isolated by affinity purification. This approach has been successfully applied to identify mRNAs [56], and more notably in the identification of lncRNA-interacting proteins that contribute to breast cancer metastasis [57]. However, one caveat of this system is the use of affinity tags that reduce specificity and limit the choice of elution buffers. A recent development has been the use of an engineered Cys4 CRISPR endoribonuclease protein that specifically binds with high affinity to a short 16-nt hairpin sequence that can be conjugated to the RNA bait via *in vitro* transcription [58]. Importantly, two point mutations in Cys4 significantly enhanced its competencies for RNP complex capture. The first mutation, Ser50 to cysteine, enables coupling to biotin, and thus immobilisation of the Cys4-RNP complex on avidin resin. The second one, His29 to alanine, inactivates RNA strand scission, which can be reversed with imidazole. This modified Cys4 thus has the ability to selectively tether RNA bait whilst enabling the release of RNA-protein complexes via inducible imidazole-mediated cleavage of the 16-nt hairpin sequence. Crucially, direct comparison of Cys4-based and biotinylated RNA affinity purification strategies revealed a dramatic decrease in the amount of background contamination to such an extent that the gel purification step after affinity purification can be omitted [58]. Although the expression of Cys4-tagged transcripts in cells has yet to be used to purify native RNPs, one can anticipate that such a method will certainly provide new insight into the composition of RNA-protein complexes, particularly in the context of lncRNA biology.

#### **Methods For Improving Stringency and Specificity During Affinity Purification:**

**Relative quantification:** Broadly speaking the relative quantification of proteins in a sample provides a wider understanding of the proteome. This is especially true within the context of interactomics analysis where relative quantification can be applied to increase confidence during the identification of protein interactions [59]. This is achieved by mixing differentially labelled lysates during affinity purification to identify specific and non-specific interactions with the tag, bait and/or affinity matrix. This labelling is achieved through marking of specific amino acids and can be performed either before collection of the sample - *metabolic labelling*, when labels are incorporated into the organism or tissue through modification of the nutrients - or at the moment of sample preparation - *chemical labelling*, in which labels are added *in vitro* to the existent protein pool. Stable isotope labelling by amino acids in cell culture (SILAC) [60] and its derivatives, such as  $^{15}\text{N}$ -labelling [61, 62] and NeuCode SILAC [63], are technologies in which cells, or even entire organisms [64, 65], are grown in media supplemented with arginine and lysine residues containing combinations of  $^{13}\text{C}$  and  $^{15}\text{N}$ , so proteins with altered masses can be generated. At the point of LC-MS, each tryptically digested peptide contains at least one of these labelled amino acids, inducing small shifts in the detected  $m/z$  ratio and enabling relative quantification of peptides from differently labelled samples, e.g. tag-alone versus tagged-bait. These SILAC-based approaches for studying protein-protein interactions have been successfully applied in a variety of fields, such as infectious diseases [66, 67] and cancer [68, 69]. Established SILAC reagents have only allowed, until recently, the analysis of three samples simultaneously, generically known as light, medium and heavy labels; however, new developments in the field have increased this number greatly, i.e. NeuCode (neutron encoded) SILAC uses a collection of isotopologues of lysine with subtle mass differences achieved by the incorporation of deuterium into its side chain, and can be used to analyse up to 18 different samples in one experiment [63, 70].

*In vitro* chemical labelling offers, in general, higher multiplicity of samples and flexibility in the design of the experimental set-up. In this process, proteins are collected and covalently linked to a specific tag that directly alters the  $m/z$  ratio of targeted peptides. Common practises include dimethyl-labelling in which free amino groups (such as side chains of lysines and the N-terminus of the peptides) react with isotopologues of cyanoborohydride and formaldehyde, generating dimethylamines, whose mass shift is detected by MS [71]. This technique is inexpensive, easy to establish and has comparable accuracy to SILAC [72]. Other tags have also been developed to generate higher levels of multiplicity. The most commonly used compounds include two derivatives of *N*-hydroxysuccinimide: tandem mass tags (TMT) [73], which use a combination of amino-reactive compounds to analyse up to 10 different samples, and isobaric tags for relative and absolute quantification (iTRAQ) [74] for up to 8 samples.

**Reducing protein complexity:** Another frequent problem in the study of protein interactions is the overall complexity of the final sample. Several approaches can be taken to focus the specificity,

frequency and localization of protein interactions including cell fractionation and labelling by proximity. The advantages of these approaches are that they reduce the potential number of hits in interactomic screens to more manageable levels. For instance using fractionation, it is possible to restrict protein populations to a particular region of the cell where the tagged or endogenous protein of interest is localised, therefore facilitating the identification of interacting partners and the reduction in background noise or non-specific interactions. The development of physical separation protocols, such as differential and gradient centrifugation, in combination with biochemical assays - that rely on the solubility of certain proteins in a variety of detergents, salt concentrations and buffers – allows the isolation of organelles or specific regions of the cell, such as lipid rafts [75] or the centrosome [76]. A method of particular relevance is protein correlation profiling (PCP) which relies on the proteomic characterization of different cellular fractions separated by gradient centrifugation and identified by immunoblotting [77]. Using these fractionation systems, the enrichment of the protein profile of a specific organelle or large protein complexes can be used to study protein-protein interactions; however, there are disadvantages including insufficient lysis, dissociation of complexes and/or re-localization of proteins to other fractions during lysis or as a result of changes in ion concentrations, pH, etc. [78, 79]. Emerging technologies to control for these limitations include localization of organelle proteins by isotope tagging (LOPIT). In this method *in vitro* chemical labelling is used to quantify the abundance of proteins along a density gradient. This pattern can then be compared to the position of reported identified proteins in the same gradient, assuming that proteins of the same organelle will localise in the same fraction [80-82].

**Proximity labelling:** Other methods use localised protein-tagging to mark and differentiate protein complexes in the cell. The success of this approach relies on the restricted modification of proteins in a specific region of the cell or proteins surrounding the bait. The most common reaction catalysed in this approach is the biotinylation of protein complexes that can be easily recovered after fractionation or stringent purification. In 2013, Rhee and colleagues used endogenous biotinylation to determine the proteome of human mitochondria [83]. In this study, ascorbate peroxidase, an enzyme that oxidises phenolic compounds, was retargeted to the inner surface of the mitochondrion to specifically biotinylate proteins in close proximity after exposure to biotin-phenol. After fractionation the mitochondria were recovered and biotinylated proteins from these organelles were purified using streptavidin beads, highlighting the power of this approach (see Figure 2).

Other recent advancements in labelling include proximity-dependent biotin identification (BioID) which allows the biotinylation of lysines in proteins surrounding a bait tagged with biotin-ligase from *E. coli* [84]. Biotinylation is triggered by incubation with biotin and labelled proteins can be stringently purified due to the strong interaction of streptavidin with biotin. This method has been used to identify interaction partners of several protein substrates as well as larger protein complexes in the context of infection [85] and cancer [86, 87]. Other techniques for labelling by proximity

include the use of cross-linking compounds combined with MS. These approaches are ideal for identifying protein interactions and studying the structure of large protein complexes [88, 89].

### **Data handling: integration, visualisation and network/pathway analysis**

There are various tools currently available for analysing the results of protein interaction studies. These are often designed with broad spectrum researchers in mind to provide an easy-to-use and accessible platform for data analysis. This review aims to provide a brief introduction to some of the powerful tools available online, or via download, but is by no means exhaustive. Comprehensive lists of available software can also be found online, e.g. at the ExPASy website [90] (also, see Table 1).

**Handling mass spectrometry data:** Raw spectra and peptide identification data from MS can be both large and complicated and it is therefore essential to establish an analysis pipeline for its handling. In most instances the initial steps in this process will be handled by a local proteomics service provider; however, as submission of raw proteomics data during publication is becoming more widely required it is advisable to have some understanding of this. A useful and expansive list of available software can be found online (see Table 1); however, some of the most popular tools include Max Quant [91] and Mascot (Matrix Science). These packages handle the identification, characterisation and quantification of peptides and are suitable for large mass-spectrometric data sets. Software is used to match identified peptides to *in silico* digested proteomes for protein identification and is also capable of quantitative analysis, i.e. label-free or stable isotope labelling in tissue culture (SILAC) technologies. Another popular piece of software for visualising the results of MS is Scaffold, from Proteome Software, which can provide information on protein function and sub-cellular localisation and also allows easy comparison between biological samples.

**Data repositories and storage:** Proteomics data in its raw format can be stored (i.e. post-publication) in the Proteomics Identifications (PRIDE) database [92]. This curated repository is hosted by the European Bioinformatics Institute (EBI) and stores detailed information such as peptide identifications, post-translational modifications (PTMs) as well as the associated spectra. Interactomics data can also be correlated to gene expression studies, e.g. microarrays; the two popular repositories for these studies are Gene Expression Omnibus (GEO) [93], part of the National Center for Biotechnology (NCBI) and EBI's Array Express [94]. These databases can be used to access published experimental data sets and both websites have useful online tools to probe gene expression levels. This cross-comparison with a user's own data can provide quick, cheap and interesting additional information on proteins of interest.

**What is protein X?** Some of the first questions about any identified protein are the most basic, e.g. what is this protein and what does it do? Manually curated databases such as the Universal Protein



Resource (Uniprot) [95], a collaboration between the EBI, Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR), are a vital tool at this point and can provide essential information on isoform sequences, post translational modifications, molecular functions, taxonomy, etc. Another useful website is the Weizmann Institute's GeneCards [96] which is very useful for identifying known aliases for any given protein.

**Identifying protein interactions and additional network analysis:** Identifying the interaction between a protein and its substrate, binding partner or small molecule is obviously an area of intense interest in modern molecular biology. As such there are a range of tools, both freeware and available under licence, to probe and map these interactions (established or predicted). Essentially these tools allow the user to examine large interactome databases, providing a broad overview of what may be known about an input and its biology. Two particularly popular online tools are IntAct [97], another service from the EBI which catalogues molecular interaction data, and String [98], a database of known and predicted protein interactions run by a cross-institute team including the University of Copenhagen, European Molecular Biology Laboratory (EMBL) and SIB. These databases include direct and indirect associations derived from a range of experimental and text-based sources and are very useful for developing a broader understanding of the molecular interactions of a protein. Also popular is the commercially available software Ingenuity Pathways Analysis (IPA) which allows subsequent integration with gene ontology networks and biological pathway analysis.

**Identifying broader biological context:** Interactomic data sets can be complicated by the number of individual proteins identified, be it from the use of crude approaches or the very nature of the baits themselves. In this instance it can be helpful to analyse the data at a more systematic level to look for enrichment of biological pathways. Two popular tools for this are Gorilla [99], a gene ontology enrichment analysis and visualisation tool, as well as the Reactome Pathway Database [100] organised by Ontario Institute for Cancer Research (OICR), New York University Medical Centre (NYUMC) and the EBI. Both tools are easily adapted to lists of proteins, rather than genes, and are designed for pathway enrichment analysis, identifying statistically over-represented gene members in input data. Reactome's online database provides an extensive library of curated pathways that can be analysed at the molecular level, providing a unique insight into the potential biological context of any identified proteins. Such tools are particularly useful when multi-partite complexes of proteins are identified rather than single binding partners.

**Visualising and modelling protein interactions at the molecular level:** A comprehensive list of available software can be found on the International Union of Crystallography's website (see Table 1); however, briefly, structural information is available for download via the Research Collaboratory for Structural Bioinformatics' (RCSB) Protein Data Bank (PDB) [101]. This powerful curated



database provides access to structural data from a range of sources and is invaluable for accessing published crystal structures and information on established interactions. Two popular packages for visualising structural data are Chimera from the University of California San Francisco (UCSF) [102] and Schrödinger's PyMOL. Both tools provide high-powered and interactive visualisation of 3D molecular structures and include features such as imaging supramolecular assemblies and results of molecular docking studies. Finally, there are also easily accessible tools for user-defined homology-modelling such as Swiss-Model [103] from SIB, which automates the process allowing easy access to a protein modelling pipeline.

### **Final Comments**

The interactomics techniques and data analysis tools described above have all successfully been applied in the laboratory setting to identify and characterise protein interactions. In most situations, they are relatively easy to establish and common enough to be available commercially. However, it is worth highlighting that many of these methodologies are now being adapted for high throughput analysis, enabling global analysis of the protein-protein, DNA-protein and RNA-protein binding landscape at an unprecedented scale. As with many emerging 'omics platforms and technologies, e.g. microarrays and next-generation sequencing, the available interactomics data 'online' now outstrips the analytical ability of most labs. As larger scale interactomics technologies come on board (e.g. affordable and robust protein microarrays) this trend will doubtlessly continue. It is therefore becoming more and more important to examine the available data online before conducting any interactomics study, as many existing large-scale studies can easily be reinterpreted without investing time in developing new systems or technical approaches.

### **Acknowledgements:**

DB and CD are both independent investigators funded by the University of Birmingham Fellowship scheme. DGC and TC are PhD students in these labs funded by the same scheme. CD is the recipient of an MRC New Investigator Award (MR/M009912/1).

### **References:**

1. McFedries, A., A. Schwaid, and A. Saghatelian, *Methods for the elucidation of protein-small molecule interactions*. Chem Biol, 2013. **20**(5): p. 667-73.
2. Silvian, L., I. Enyedy, and G. Kumaravel, *Inhibitors of protein-protein interactions: new methodologies to tackle this challenge*. Drug Discov Today Technol, 2013. **10**(4): p. e509-15.
3. Arkin, M.R., Y. Tang, and J.A. Wells, *Small-molecule inhibitors of protein-protein interactions: progressing toward the reality*. Chem Biol, 2014. **21**(9): p. 1102-14.
4. Fields, S. and O. Song, *A novel genetic system to detect protein-protein interactions*. Nature, 1989. **340**(6230): p. 245-6.
5. Brückner, A., et al., *Yeast Two-Hybrid, a Powerful Tool for Systems Biology*. International Journal of Molecular Sciences, 2009. **10**(6): p. 2763-2788.

6. Ehrhard, K.N., et al., *Use of G-protein fusions to monitor integral membrane protein-protein interactions in yeast*. Nat Biotechnol, 2000. **18**(10): p. 1075-9.
7. Dirnberger, D., M. Messerschmid, and R. Baumeister, *An optimized split-ubiquitin cDNA-library screening system to identify novel interactors of the human Frizzled 1 receptor*. Nucleic Acids Res, 2008. **36**(6): p. e37.
8. Johnsson, N. and A. Varshavsky, *Split ubiquitin as a sensor of protein interactions in vivo*. Proc Natl Acad Sci U S A, 1994. **91**(22): p. 10340-4.
9. Urech, D.M., P. Lichtlen, and A. Barberis, *Cell growth selection system to detect extracellular and transmembrane protein interactions*. Biochim Biophys Acta, 2003. **1622**(2): p. 117-27.
10. Serebriiskii, I., V. Khazak, and E.A. Golemis, *A two-hybrid dual bait system to discriminate specificity of protein interactions*. J Biol Chem, 1999. **274**(24): p. 17080-7.
11. Koegl, M. and P. Uetz, *Improving yeast two-hybrid screening systems*. Briefings in Functional Genomics & Proteomics, 2007. **6**(4): p. 302-312.
12. Iyer, K., et al., *Utilizing the split-ubiquitin membrane yeast two-hybrid system to identify protein-protein interactions of integral membrane proteins*. Sci STKE, 2005. **2005**(275): p. pl3.
13. Paumi, C.M., et al., *Mapping protein-protein interactions for the yeast ABC transporter Ycf1p by integrated split-ubiquitin membrane yeast two-hybrid analysis*. Mol Cell, 2007. **26**(1): p. 15-25.
14. Petschnigg, J., et al., *The mammalian-membrane two-hybrid assay (MaMTH) for probing membrane-protein interactions in human cells*. Nat Methods, 2014. **11**(5): p. 585-92.
15. Kwon, J.A., et al., *High diagnostic accuracy of antigen microarray for sensitive detection of hepatitis C virus infection*. Clin Chem, 2008. **54**(2): p. 424-8.
16. Zhu, H., et al., *Analysis of yeast protein kinases using protein chips*. Nat Genet, 2000. **26**(3): p. 283-9.
17. Zhu, H., et al., *Global Analysis of Protein Activities Using Proteome Chips*. Science, 2001. **293**(5537): p. 2101-2105.
18. Uzoma, I. and H. Zhu, *Interactome mapping: using protein microarray technology to reconstruct diverse protein networks*. Genomics, proteomics & bioinformatics, 2013. **11**(1): p. 18-28.
19. Zhu, J., et al., *RNA-binding proteins that inhibit RNA virus infection*. Proc Natl Acad Sci U S A, 2007. **104**(9): p. 3129-34.
20. Berger, M.F. and M.L. Bulyk, *Universal protein binding microarrays for the comprehensive characterization of the DNA binding specificities of transcription factors*. Nat Protoc, 2009. **4**(3): p. 393-411.
21. Ho, S.W., et al., *Linking DNA-binding proteins to their recognition sequences by using protein microarrays*. Proc Natl Acad Sci U S A, 2006. **103**(26): p. 9940-5.
22. Yates, J.R., C.I. Ruse, and A. Nakorchevsky, *Proteomics by mass spectrometry: approaches, advances, and applications*. Annu Rev Biomed Eng, 2009. **11**: p. 49-79.
23. Cargile, B.J., J.L. Bundy, and J.L. Stephenson, Jr., *Potential for false positive identifications from large databases through tandem mass spectrometry*. J Proteome Res, 2004. **3**(5): p. 1082-5.
24. Kumar, C. and M. Mann, *Bioinformatics analysis of mass spectrometry-based proteomics data sets*. FEBS Letters, 2009. **583**(11): p. 1703-1712.
25. Jäger, S., et al., *Global landscape of HIV-human protein complexes*. Nature, 2012. **481**(7381): p. 365-370.
26. Krogan, N.J., et al., *Global landscape of protein complexes in the yeast Saccharomyces cerevisiae*. Nature, 2006. **440**(7084): p. 637-43.
27. Rigaut, G., et al., *A generic protein purification method for protein complex characterization and proteome exploration*. Nat Biotechnol, 1999. **17**(10): p. 1030-2.
28. Bailey, D., et al., *Identification of Protein Interacting Partners Using Tandem Affinity Purification*. Journal of Visualized Experiments : JoVE, 2012(60): p. 3643.
29. Goodfellow, I. and D. Bailey, *Detection of protein-protein interactions using tandem affinity purification*. Methods Mol Biol, 2014. **1177**: p. 121-33.
30. Puig, O., et al., *The tandem affinity purification (TAP) method: a general procedure of protein complex purification*. Methods, 2001. **24**(3): p. 218-29.

31. Gunzl, A. and B. Schimanski, *Tandem affinity purification of proteins*. Curr Protoc Protein Sci, 2009. **Chapter 19**: p. Unit 19 19.
32. Chung, L., et al., *Norovirus Translation Requires an Interaction between the C Terminus of the Genome-linked Viral Protein VPg and Eukaryotic Translation Initiation Factor 4G*. The Journal of Biological Chemistry, 2014. **289**(31): p. 21738-21750.
33. Haura, E.B., et al., *Using iTRAQ® Combined with Tandem Affinity Purification to Enhance Low-abundance Proteins Associated with Somatic-mutated EGFR Core Complexes in Lung Cancer*. Journal of proteome research, 2011. **10**(1): p. 182-190.
34. Hubner, N.C., et al., *Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions*. J Cell Biol, 2010. **189**(4): p. 739-54.
35. Poser, I., et al., *BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals*. Nature methods, 2008. **5**(5): p. 409-415.
36. Salic, A. and T.J. Mitchison, *A chemical method for fast and sensitive detection of DNA synthesis in vivo*. Proc Natl Acad Sci U S A, 2008. **105**(7): p. 2415-20.
37. Sirbu, B.M., et al., *Analysis of protein dynamics at active, stalled, and collapsed replication forks*. Genes Dev, 2011. **25**(12): p. 1320-7.
38. Sirbu, B.M., et al., *Identification of proteins at active, stalled, and collapsed replication forks using isolation of proteins on nascent DNA (iPOND) coupled with mass spectrometry*. J Biol Chem, 2013. **288**(44): p. 31458-67.
39. Gillespie, P.J., A. Gambus, and J.J. Blow, *Preparation and use of Xenopus egg extracts to study DNA replication and chromatin associated proteins*. Methods, 2012. **57**(2): p. 203-13.
40. Cross, M.K. and M.A. Powers, *Learning about cancer from frogs: analysis of mitotic spindles in Xenopus egg extracts*. Dis Model Mech, 2009. **2**(11-12): p. 541-7.
41. Masui, Y. and C.L. Markert, *Cytoplasmic control of nuclear behavior during meiotic maturation of frog oocytes*. J Exp Zool, 1971. **177**(2): p. 129-45.
42. Ma, L., et al., *Identification of small ubiquitin-like modifier substrates with diverse functions using the Xenopus egg extract system*. Mol Cell Proteomics, 2014. **13**(7): p. 1659-75.
43. Leach, C.A. and W.M. Michael, *Ubiquitin/SUMO modification of PCNA promotes replication fork progression in Xenopus laevis egg extracts*. J Cell Biol, 2005. **171**(6): p. 947-54.
44. Moreno, S.P., et al., *Polyubiquitylation drives replisome disassembly at the termination of DNA replication*. Science, 2014. **346**(6208): p. 477-81.
45. Murat, P. and S. Balasubramanian, *Existence and consequences of G-quadruplex structures in DNA*. Curr Opin Genet Dev, 2014. **25**: p. 22-9.
46. Huppert, J.L., *Structure, location and interactions of G-quadruplexes*. FEBS J, 2010. **277**(17): p. 3452-8.
47. von Hacht, A., et al., *Identification and characterization of RNA guanine-quadruplex binding proteins*. Nucleic Acids Res, 2014. **42**(10): p. 6630-44.
48. Yang, G., X. Lu, and L. Yuan, *LncRNA: a link between RNA and cancer*. Biochimica et biophysica acta, 2014. **1839**(11): p. 1097-109.
49. Keene, J.D., J.M. Komisarow, and M.B. Friedersdorf, *RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts*. Nature protocols, 2006. **1**(1): p. 302-7.
50. Mili, S. and J.A. Steitz, *Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses*. RNA, 2004. **10**(11): p. 1692-4.
51. Ule, J., et al., *CLIP identifies Nova-regulated RNA networks in the brain*. Science, 2003. **302**(5648): p. 1212-5.
52. Hafner, M., et al., *Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP*. Cell, 2010. **141**(1): p. 129-41.
53. Andersen, C.B.F., et al., *Structure of the exon junction core complex with a trapped DEAD-box ATPase bound to RNA*. Science, 2006. **313**(5795): p. 1968-1972.
54. Singh, G., E.P. Ricci, and M.J. Moore, *RIPit-Seq: a high-throughput approach for footprinting RNA:protein complexes*. Methods, 2014. **65**(3): p. 320-32.

55. Beach, D.L. and J.D. Keene, *Ribotrap : targeted purification of RNA-specific RNPs from cell lysates through immunoaffinity precipitation to identify regulatory proteins and RNAs*. Methods in molecular biology, 2008. **419**: p. 69-91.
56. Yoon, J.H., S. Srikantan, and M. Gorospe, *MS2-TRAP (MS2-tagged RNA affinity purification): tagging RNA to identify associated miRNAs*. Methods, 2012. **58**(2): p. 81-7.
57. Gumireddy, K., et al., *Identification of a long non-coding RNA-associated RNP complex regulating metastasis at the translational step*. Embo Journal, 2013. **32**(20): p. 2672-2684.
58. Lee, H.Y., et al., *RNA-protein analysis using a conditional CRISPR nuclease*. Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(14): p. 5416-21.
59. Mann, M., *Functional and quantitative proteomics using SILAC*. Nat Rev Mol Cell Biol, 2006. **7**(12): p. 952-8.
60. Ong, S.E., et al., *Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics*. Mol Cell Proteomics, 2002. **1**(5): p. 376-86.
61. Ippel, J.H., et al., *In vivo uniform (15)N-isotope labelling of plants: using the greenhouse for structural proteomics*. Proteomics, 2004. **4**(1): p. 226-34.
62. Washburn, M.P., et al., *Analysis of Quantitative Proteomic Data Generated via Multidimensional Protein Identification Technology*. Analytical Chemistry, 2002. **74**(7): p. 1650-1657.
63. Hebert, A.S., et al., *Neutron-encoded mass signatures for multiplexed proteome quantification*. Nat Methods, 2013. **10**(4): p. 332-4.
64. Kruger, M., et al., *SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function*. Cell, 2008. **134**(2): p. 353-64.
65. Larance, M., et al., *Stable Isotope Labeling with Amino acids in Nematodes*. Nature methods, 2011. **8**(10): p. 849-851.
66. Munday, D.C., et al., *Using SILAC and quantitative proteomics to investigate the interactions between viral and host proteomes*. Proteomics, 2012. **12**(4-5): p. 666-72.
67. Barrero, C.A., et al., *HIV-1 Vpr Modulates Macrophage Metabolic Pathways: A SILAC-Based Quantitative Analysis*. PLoS ONE, 2013. **8**(7): p. e68376.
68. Geiger, T., et al., *Super-SILAC mix for quantitative proteomics of human tumor tissue*. Nat Methods, 2010. **7**(5): p. 383-5.
69. Zanivan, S., et al., *In Vivo SILAC-Based Proteomics Reveals Phosphoproteome Changes during Mouse Skin Carcinogenesis*. Cell Reports. **3**(2): p. 552-566.
70. Rose, C.M., et al., *Neutron encoded labeling for peptide identification*. Anal Chem, 2013. **85**(10): p. 5129-37.
71. Boersema, P.J., et al., *Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics*. Nat Protoc, 2009. **4**(4): p. 484-94.
72. Lau, H.T., et al., *Comparing SILAC- and stable isotope dimethyl-labeling approaches for quantitative proteomics*. J Proteome Res, 2014. **13**(9): p. 4164-74.
73. Thompson, A., et al., *Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS*. Anal Chem, 2003. **75**(8): p. 1895-904.
74. Ross, P.L., et al., *Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents*. Mol Cell Proteomics, 2004. **3**(12): p. 1154-69.
75. Foster, L.J., C.L. De Hoog, and M. Mann, *Unbiased quantitative proteomics of lipid rafts reveals high specificity for signaling factors*. Proc Natl Acad Sci U S A, 2003. **100**(10): p. 5813-8.
76. Andersen, J.S., et al., *Proteomic characterization of the human centrosome by protein correlation profiling*. Nature, 2003. **426**(6966): p. 570-4.
77. Foster, L.J., et al., *A mammalian organelle map by protein correlation profiling*. Cell, 2006. **125**(1): p. 187-99.
78. Huber, L.A., K. Pfaller, and I. Vietor, *Organelle proteomics: implications for subcellular fractionation in proteomics*. Circ Res, 2003. **92**(9): p. 962-8.
79. Tinnikov, A.A. and H.H. Samuels, *A novel cell lysis approach reveals that caspase-2 rapidly translocates from the nucleus to the cytoplasm in response to apoptotic stimuli*. PLoS One, 2013. **8**(4): p. e61085.



80. Dunkley, T.P., et al., *Localization of organelle proteins by isotope tagging (LOPIT)*. Mol Cell Proteomics, 2004. **3**(11): p. 1128-34.
81. Dunkley, T.P., et al., *Mapping the Arabidopsis organelle proteome*. Proc Natl Acad Sci U S A, 2006. **103**(17): p. 6518-23.
82. Christoforou, A., A.M. Arias, and K.S. Lilley, *Determining protein subcellular localization in mammalian cell culture with biochemical fractionation and iTRAQ 8-plex quantification*. Methods Mol Biol, 2014. **1156**: p. 157-74.
83. Rhee, H.-W., et al., *Proteomic Mapping of Mitochondria in Living Cells via Spatially-Restricted Enzymatic Tagging*. Science (New York, N.Y.), 2013. **339**(6125): p. 1328-1331.
84. Roux, K.J., et al., *A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells*. J Cell Biol, 2012. **196**(6): p. 801-10.
85. Chen, A.L., et al., *Novel components of the Toxoplasma inner membrane complex revealed by BioID*. MBio, 2015. **6**(1): p. e02357-14.
86. Coyaud, E., et al., *BioID-based identification of SCF beta-TrCP1/2 E3 ligase substrates*. Mol Cell Proteomics, 2015.
87. Dingar, D., et al., *BioID identifies novel c-MYC interacting partners in cultured cells and xenograft tumors*. J Proteomics, 2015. **118**: p. 95-111.
88. Schmidt, C., et al., *Comparative cross-linking and mass spectrometry of an intact F-type ATPase suggest a role for phosphorylation*. Nat Commun, 2013. **4**.
89. Liu, F., et al., *An enhanced protein crosslink identification strategy using CID-cleavable chemical crosslinkers and LC/MS(n) analysis*. Proteomics, 2012. **12**(3): p. 401-5.
90. Artimo, P., et al., *ExPASy: SIB bioinformatics resource portal*. Nucleic Acids Res, 2012. **40**(Web Server issue): p. W597-603.
91. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification*. Nat Biotechnol, 2008. **26**(12): p. 1367-72.
92. Vizcaino, J.A., et al., *The PRoteomics IDentifications (PRIDE) database and associated tools: status in 2013*. Nucleic Acids Res, 2013. **41**(Database issue): p. D1063-9.
93. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets--update*. Nucleic Acids Res, 2013. **41**(Database issue): p. D991-5.
94. Kolesnikov, N., et al., *ArrayExpress update--simplifying data submissions*. Nucleic Acids Res, 2015. **43**(Database issue): p. D1113-6.
95. UniProt, C., *UniProt: a hub for protein information*. Nucleic Acids Res, 2015. **43**(Database issue): p. D204-12.
96. Safran, M., et al., *GeneCards Version 3: the human gene integrator*. Database (Oxford), 2010. **2010**: p. baq020.
97. Orchard, S., et al., *The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases*. Nucleic Acids Res, 2014. **42**(Database issue): p. D358-63.
98. Szklarczyk, D., et al., *STRING v10: protein-protein interaction networks, integrated over the tree of life*. Nucleic Acids Res, 2015. **43**(Database issue): p. D447-52.
99. Eden, E., et al., *GORilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists*. BMC Bioinformatics, 2009. **10**: p. 48.
100. Croft, D., et al., *Reactome: a database of reactions, pathways and biological processes*. Nucleic Acids Res, 2011. **39**(Database issue): p. D691-7.
101. Berman, H.M., et al., *The Protein Data Bank*. Nucleic Acids Res, 2000. **28**(1): p. 235-42.
102. Pettersen, E.F., et al., *UCSF Chimera--a visualization system for exploratory research and analysis*. J Comput Chem, 2004. **25**(13): p. 1605-12.
103. Arnold, K., et al., *The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling*. Bioinformatics, 2006. **22**(2): p. 195-201.

# Figure Legends:

**Figure 1: Affinity purification of protein complexes.** Experimental design for interactomics experiments can be improved through the application of novel techniques such as protein correlation profiling and endogenous biotinylation; methods that can reduce the complexity of samples to be analysed by mass spectrometry. Various techniques for affinity purification of protein complexes including tandem affinity purification and RNA-protein immunoprecipitation in tandem can then be applied to identify protein complexes, although this is dependent on the type of bait used (protein, DNA or RNA). Finally, quantification of proteins can also be applied to reduce complexity, increase stringency and also to provide statistical confidence that identified proteins are bait-specific.

**Figure 2: Schematic representation of spatially-restricted endogenous enzymatic tagging.** The ascorbate peroxidase (APEX) protein can be retargeted to different sub-cellular locations in the cell by the addition of organelle specific signal peptides, e.g. mitochondrial localisation signals. Re-engineered APEX expression and localisation can then be determined by microscopy. Enzymatic tagging proceeds after addition of biotin-phenol and  $H_2O_2$  to cells for a short period (1 minute). During this incubation the enzymatic activity of APEX produces short-lived biotin-phenolic radicals that react only with proteins in close proximity, tagging them with biotin. After cell lysis, biotinylated protein complexes can be purified using high stringency streptavidin beads and subsequently analysed by mass spectrometry.



684 **Table 1:** Online resources and software for interactomics researchers.

Resource	Description	URL	Reference
Swiss Institute of Bioinformatics Expert Protein Analysis System	Software and program access/list	<a href="http://www.expasy.org/">http://www.expasy.org/</a>	Artimo <i>et al.</i> , 2012
Molecular biology tools	List of online (often freeware) analysis tools	<a href="http://molbiol-tools.ca/">http://molbiol-tools.ca/</a>	n/a software and program list, Dr. Andrew Kropinski
PRIDE Proteomics Identifications	Proteomics data repository	<a href="http://www.ebi.ac.uk/pride/archive/">http://www.ebi.ac.uk/pride/archive/</a>	Vizcaino <i>et al.</i> , 2013
National Center for Biotechnology Gene Expression Omnibus (GEO)	Gene expression experimental data repository	<a href="http://www.ncbi.nlm.nih.gov/geo/">http://www.ncbi.nlm.nih.gov/geo/</a>	Barrett <i>et al.</i> , 2013
European Bioinformatics Institute Array Express	Archive of functional genomics data	<a href="http://www.ebi.ac.uk/arrayexpress/">http://www.ebi.ac.uk/arrayexpress/</a>	Kolesnikov <i>et al.</i> , 2015
IntAct	Protein interactions database	<a href="http://www.ebi.ac.uk/intact/">http://www.ebi.ac.uk/intact/</a>	Orchard <i>et al.</i> , 2014
String	Protein interactions database	<a href="http://string-db.org/">http://string-db.org/</a>	Szklarczyk <i>et al.</i> , 2015
Ingenuity Pathways Analysis (IPA)	Commercially available biological pathway analysis software	<a href="http://www.ingenuity.com/products/ipa">http://www.ingenuity.com/products/ipa</a>	n/a, commercial software
Mass spectrometry utilities	Software and program list for handling mass spectrometry data	<a href="http://www.ms-utils.org/wiki/pmwiki.php/Main/SoftwareList">http://www.ms-utils.org/wiki/pmwiki.php/Main/SoftwareList</a>	n/a, software and program list
Matrix Science Mascot	Software to process mass spectrometry data	<a href="http://www.matrixscience.com/index.html">http://www.matrixscience.com/index.html</a>	n/a, commercial software
Max Quant	Software to process mass spectrometry data	<a href="http://141.61.102.17/maxquant_doku/doku.php?id=start">http://141.61.102.17/maxquant_doku/doku.php?id=start</a>	Cox & Mann, 2008
Scaffold	Software to handle mass spectrometry data	<a href="http://www.proteomesoftware.com/products/scaffold/">http://www.proteomesoftware.com/products/scaffold/</a>	n/a, commercial software
The Universal Protein Resource (Uniprot)	Curated database of protein information	<a href="http://www.uniprot.org/">http://www.uniprot.org/</a>	Uniprot Consortium, 2015
Weizmann Institute's GeneCards	Database of genetic information	<a href="http://www.genecards.org/">http://www.genecards.org/</a>	Safran <i>et al.</i> , 2010
Reactome	Online tool for analysing biological pathways	<a href="http://www.reactome.org/">http://www.reactome.org/</a>	Croft <i>et al.</i> , 2011
Gorilla: Gene Ontology enrichment analysis and visualization tool	Online tool for identifying enriched pathways in data sets	<a href="http://cbl-gorilla.cs.technion.ac.il/">http://cbl-gorilla.cs.technion.ac.il/</a>	Eden <i>et al.</i> , 2009
International Union of Crystallography Resources Page	Extensive list of structural software	<a href="http://www.iucr.org/resources/other-directories/software">http://www.iucr.org/resources/other-directories/software</a>	n/a, software and program list
RCSB PDB database	Database of protein structures	<a href="http://www.rcsb.org/pdb/home/home.do">http://www.rcsb.org/pdb/home/home.do</a>	Berman <i>et al.</i> , 2000
University of California San Francisco's Chimera	Software for analysing structural data	<a href="http://www.cgl.ucsf.edu/chimera/">http://www.cgl.ucsf.edu/chimera/</a>	Pettersen <i>et al.</i> , 2004
PyMOL	Software for analysing structural data	<a href="https://www.pymol.org/">https://www.pymol.org/</a>	n/a, commercial software
SIB Swiss-Model	Online tool for protein modelling	<a href="http://swissmodel.expasy.org/">http://swissmodel.expasy.org/</a>	Arnold <i>et al.</i> , 2006

685

686

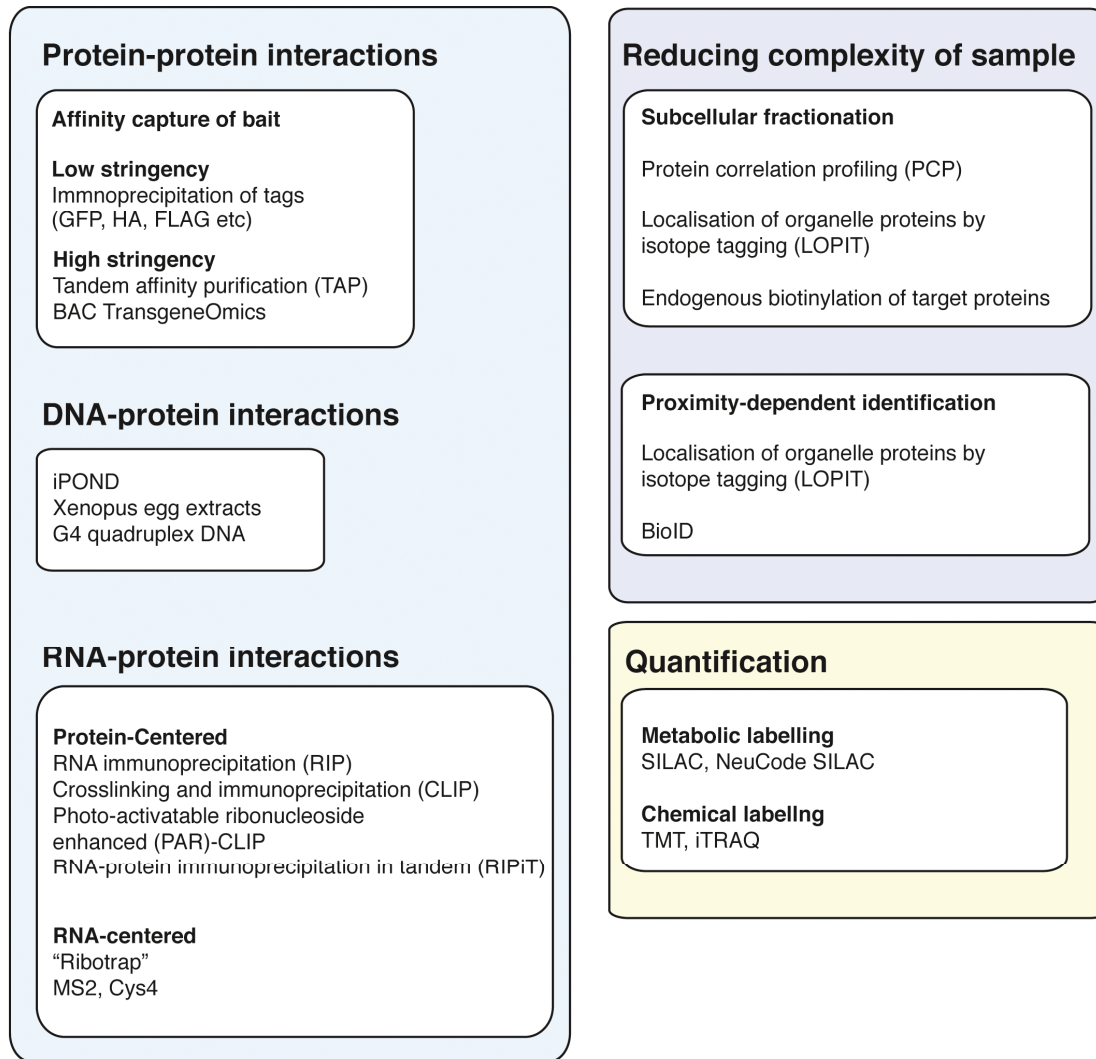


Figure 1: Gonçalves Carneiro et al.

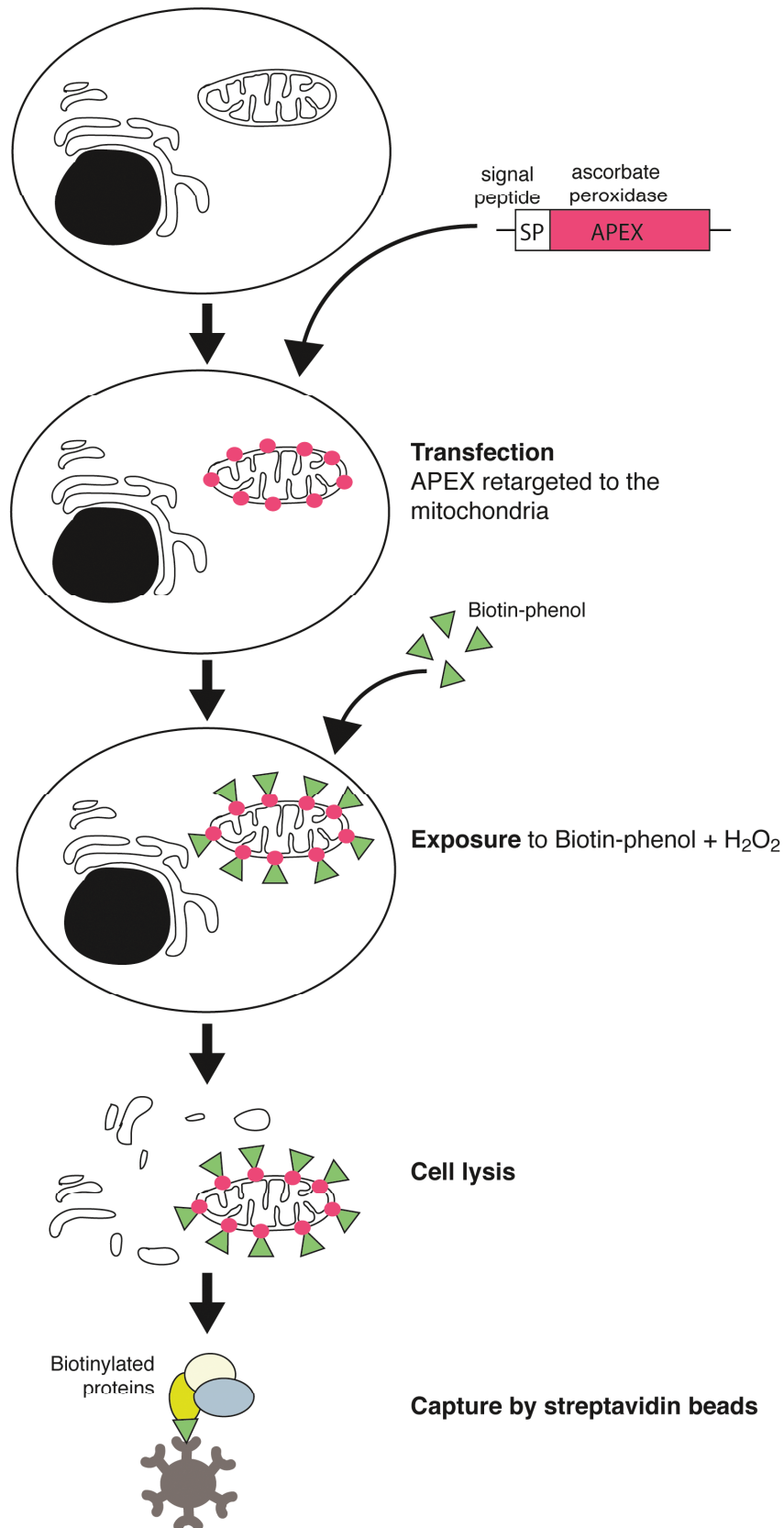


Figure 2: Gonçalves Carneiro et al.

690

691 Highlights:

- 692       **1. Introducing interactomics techniques:** We have provided an introductory review of a broad range of  
693 methodologies used to identify protein interactions, particularly emerging technologies in the fields  
694 of protein-protein, DNA-protein and RNA-protein interactomics.
- 695       **2. Improving experimental set-up:** We describe some of the powerful approaches researchers can  
696 apply to interactomics to improve stringency and specificity in their experiments. These include  
697 quantitative labelling, proximity labelling and sub-cellular fractionation.
- 698       **3. Downstream processing:** There are numerous online tools and freeware for interrogating and  
699 storing interactomics data. We have provided an overview of these together with links and  
700 references.

701